The Heckman correction: an introduction

Jimmy Jin

July 22, 2016

The basic setup

You have a population of N individuals and want to fit a linear model to this population:

$$y = \boldsymbol{X}\boldsymbol{\beta} + \epsilon \tag{1}$$

You have a sample of size n from this population but there's a problem: it was not selected randomly, and so is probably not representative of the population. Clearly if you ran a regression using this sample, you might be biased estimates. The Heckman correction is designed to fix this.

Begin by modeling whether or not an individual in the population will be selected into the sample. Let s be a binary variable representing this sample selection, so that if $s_i = 1$ then the ith individual is sampled. The model that the Heckman correction is based on is another linear model whose covariates are generally a superset of the original design matrix $\boldsymbol{X}$:

$$s = \mathbb{1}\{\tilde{\boldsymbol{X}}\boldsymbol{\delta} + \tau > 0\} \tag{2}$$

1. $\tilde{\boldsymbol{X}}$ is a design matrix; a superset of the base design matrix $\boldsymbol{X}$
2. $\boldsymbol{\delta}$ is a vector of coefficients for the model
3. τ is a noise term for the linear model

The assumptions

For the Heckman correction to produce consistent estimates, a couple assumptions have to be made.

1. (ϵ, τ) is independent of $\tilde{\boldsymbol{X}}$ and is mean zero, i.e. $\mathbb{E}((\epsilon, \tau)) = (0, 0)$
2. $\tau \sim$ Normal(0,1)
3. $\mathbb{E}(\epsilon|\tau) = \gamma\tau$ for some $\gamma \in \mathbb{R}$

The third is key. It basically says that the two error terms are linearly related, and also specifies a parameter γ which turns out to control the degree to which the sample selection biases estimation of $\boldsymbol{\beta}$.

Derivation of the bias (skippable)

In practice, since all we observe is the subset of the population for which $s = 1$, we want to know what the relationship is between $y|\boldsymbol{X}, s$ and $\boldsymbol{\beta}$. That will show us the right specification of model (1) to eliminate bias from sample selection.

First we first need to calculate what we might expect out of y if we knew the selection design matrix $\tilde{\boldsymbol{X}}$ and also the value of the error term τ of the selection model:

$$\begin{aligned}\mathbb{E}(y|\tilde{\boldsymbol{X}}, \tau) &= \boldsymbol{X}\boldsymbol{\beta} + \mathbb{E}(\epsilon|\tilde{\boldsymbol{X}}, \tau) \\ &= \boldsymbol{X}\boldsymbol{\beta} + \mathbb{E}(\epsilon|\tau) \\ &= \boldsymbol{X}\boldsymbol{\beta} + \gamma\tau\end{aligned}$$

Going from the first to the second line we used the assumption (ϵ, τ) are independent of $\tilde{\boldsymbol{X}}$. In all, this tells us that given information about $\tilde{\boldsymbol{X}}$ and the noise in the selection model, the bias we can expect is $\gamma\tau$.

Note here that this tells us something more. The bias is equal to $\gamma\tau$. If γ is zero, then the bias term drops out entirely. More generally in the case where we don't assume that $\mathbb{E}(\epsilon|\tau) = \gamma\tau$, we see that a sufficient condition for no bias is for γ to be independent of τ. This is interesting because $\tilde{\boldsymbol{X}}$ can include more covariates than $\boldsymbol{X}$. But because we aren't trying to estimate the coefficients of those covariates, it does not bias our results.

Now in practice what we really observe is not the error term of the selection model τ but rather just the coarser view of s (i.e. whether or not the point appeared in our sample). So calculating the expectation of our response y conditioning on s instead, we get

$$\begin{aligned}\mathbb{E}(y|\tilde{\boldsymbol{X}}, s) &= \mathbb{E}\Big[\mathbb{E}(y|\tilde{\boldsymbol{X}}, \tau)|\tilde{\boldsymbol{X}}, s\Big] \\ &= \mathbb{E}\Big[\boldsymbol{X}\boldsymbol{\beta} + \gamma\tau|\tilde{\boldsymbol{X}}, s\Big] \\ &= \boldsymbol{X}\boldsymbol{\beta} + \gamma\mathbb{E}(\tau|\tilde{\boldsymbol{X}}, s)\end{aligned}$$

The in-sample bias and the inverse Mills' ratio

The last line of math above tells us that the bias introduced to the model by non-random sample selection is equal to $\gamma\mathbb{E}(\tau|\tilde{\boldsymbol{X}}, s)$.

Now since $s = 1$ corresponds to the case of the data we actually observe in our sample, we can further simplify the bias expression by calculating $\mathbb{E}(\tau|\tilde{\boldsymbol{X}}, s = 1)$. Recall that by (2), if we know the values of $\tilde{\boldsymbol{X}}$ and $s = 1$, then it must be true that

$$\tau > -\tilde{\boldsymbol{X}}\boldsymbol{\delta}$$

Furthermore since we assume that $\tau \sim N(0, 1)$, then conditional on $s = 1$, τ must follow a truncated standard Normal distribution with truncation at $-\tilde{\boldsymbol{X}}\boldsymbol{\delta}$. It follows then that the expected value of τ is

$$\mathbb{E}(\tau|\tilde{\boldsymbol{X}}, s = 1) = \frac{\phi(\tilde{\boldsymbol{X}}\boldsymbol{\delta})}{\Phi(\tilde{\boldsymbol{X}}\boldsymbol{\delta})} := \lambda(\tilde{\boldsymbol{X}}\boldsymbol{\delta})$$

where $\lambda(\cdot)$ is the *inverse Mills' ratio*.

To summarize, the Mills' ratio arises **only because** of the linear component inside our selection model. Because sample selection is modelled as whether or not a linear component ($\tilde{\boldsymbol{X}}\boldsymbol{\delta}$) exceeds a threshold, then conditioning on the sample selection = 1 is just imposing a requirement on the linear component being above that threshold.

Estimation

Now we know that the true mean of the biased-sample observations is

$$\mathbb{E}(y|\tilde{\boldsymbol{X}}, s = 1) = \boldsymbol{X}\boldsymbol{\beta} + \gamma\lambda(\tilde{\boldsymbol{X}}\boldsymbol{\delta})$$

This means that if we just knew $\lambda(\tilde{\boldsymbol{X}}\boldsymbol{\delta})$, then we could throw those into a regression and get an **unbiased** estimate of β (as well as an unbiased estimate of γ).

One problem: we don't know $\boldsymbol{\delta}$. But we can estimate it. We have data on observations that are all inside ($s = 1$) and outside ($s = 0$) the sample of interest, as well as the covariates $\tilde{\boldsymbol{x}}$ for all those points. We also know that the covariates are mapped to $\{0, 1\}$ after being subjected to standard normal noise. Therefore it makes sense to estimate $\boldsymbol{\delta}$ by probit regression.

The final procedure therefore is

1. Estimate $\hat{\boldsymbol{\delta}}$ from the probit model

   $$\mathbb{P}(s = 1|\tilde{\boldsymbol{X}}) = \Phi(\tilde{\boldsymbol{X}}\hat{\boldsymbol{\delta}})$$

   using data on all N individuals in the population.

2. Use $\hat{\boldsymbol{\delta}}$ to estimate the inverse Mills' ratios for each point i in your subsample $\lambda(\tilde{\boldsymbol{X}}_{\boldsymbol{i}}\hat{\boldsymbol{\delta}})$.

3. Fit the base model on your subsample with the estimated inverse Mills' ratios added

   $$y_i = \boldsymbol{X}_{\boldsymbol{i}}\boldsymbol{\beta} + \lambda(\tilde{\boldsymbol{X}}_{\boldsymbol{i}}\hat{\boldsymbol{\delta}}) + \epsilon_i$$

There are two things to note:

1. **Interpretation of the first step:** The first step is "sort of" estimating the probability that an individual will be in the sample, but not quite. Instead, it's estimating the parameter $\boldsymbol{\delta}$ which *in turn* controls the selection model (2) that determines whether or not an individual enters the sample.

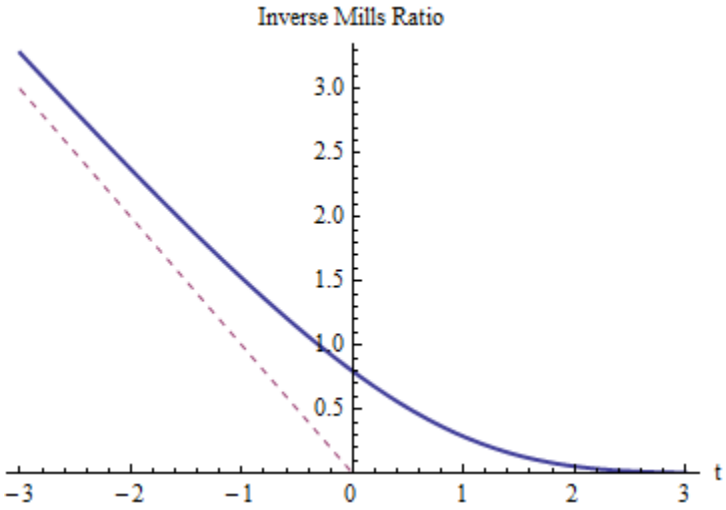
2. **Consistency:** Since we are using $\hat{\boldsymbol{\delta}}$ instead of the true $\boldsymbol{\delta}$, it is no longer guaranteed that our estimates are unbiased. However, it is shown in Heckman's original paper that the estimates $\hat{\boldsymbol{\beta}}$ are consistent.

The issue of the exclusion restriction

Nothing above relies on the presence of an exclusion restriction. In other words, the data matrices $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ can be exactly the same. So long as the design matrices in

(1) and (2) are of full rank, then the Heckman correction will produce consistent estimates.

However, there is a potential issue with multicollinearity. The inverse Mills' ratio $\lambda(\cdot)$ is very close to linear over a large portion of its range



Thus, if $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ contain exactly the same covariates and the range in $\tilde{\boldsymbol{X}}$ is not that large, it is possible that the $\lambda(\tilde{\boldsymbol{X}}_{\boldsymbol{i}})$'s will behave just like linear functions of the $\tilde{\boldsymbol{X}}_{\boldsymbol{i}}$'s. Then inserting these back into the second step will cause severe multicollinearity and large standard errors.

One way to get around this is to impose exclusion restrictions on $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$—in other words, throw in covariates into $\tilde{\boldsymbol{X}}$ which are not in $\boldsymbol{X}$ or highly correlated with existing covariates. But at the end of the day, it isn't required to perform the correction itself.

References

[1] James J. Heckman, *Sample Selection Bias as a Specification Error*. Econometrica, 47(1):153-161, 1979.

[2] Jeffrey M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Cambridge, Massachusetts, 2010.